Varieties of Plain Language

Allen Riddell
Indiana University Bloomington
Bloomington, Indiana, USA
riddella@indiana.edu

Yohei Igarashi
University of Connecticut
Storrs, Connecticut, USA
yohei.igarashi@uconn.edu

Abstract

Many organizations seek or need to produce documents that are written plainly. In the United States, the "Plain Writing Act of 2010" requires that many federal agencies' documents for the public are written in plain English. In particular, the government's Plain Language Action and Information Network ("PLAIN") recommends that writers use short sentences and everyday words, as does the Securities and Exchange Commission's "Plain English Rule." Since the 1970s, American plain language advocates have moved away from readability measures and favored usability testing and document design considerations. But in this paper we use quantitative measures of sentence length and word difficulty that (1) reveal stylistic variation among PLAIN's exemplars of plain writing, and (2) help us position PLAIN's exemplars relative to documents written in other kinds of accessible English (e.g., *The New York Times*, Voice of America Special English, and *Wikipedia*) and one academic document likely to be perceived as difficult. Uncombined measures for sentences and vocabulary—left separate, unlike in traditional readability formulas—can complement usability testing and document design considerations, and advance knowledge about different types of plainer English.

1 Introduction

The quality of being "plain" has been held up as a stylistic ideal in English prose since the later seventeenth century (Guillory, 2017). This ideal has shown remarkable persistence (Cutts, 2020). In the United States, the plain language movement took off in the 1940s, and plainness remains a stylistic goal for many kinds of organizations in their writing on websites and other publications: medical and public health information, insurance policies, instructions to jurors, loan agreements, and Social Security benefits statements, to name just a few (Schriver, 2017; Cutts, 2020).

Since the passage of the Plain Writing Act of 2010, American federal agencies must use "plain language," defined as "writing that is clear, concise, well-organized, and follows other best practices appropriate to the subject or field and intended audience" (United States Congress, 2010). Affecting all 2.1 million employees of the U.S. Federal government (Jennings and Nagel, 2020), the Act requires that agencies use this kind of language in many of their documents for the public, train their employees in this style, and demonstrate their compliance with the Act.[1] There are several cogent rationales for plain language use: it grants access to understandable information to a greater number of people with differing literacy levels; it saves agencies the labor and money involved in clearing up confusing communications; and it may help to restore citizens' trust in public-serving organizations, an especially important agenda in our era of misinformation, disinformation, and propaganda (Schriver, 2017). If we add to the U.S. government's plain writing imperative the similarly simple writing style espoused by Big Tech in all of its apps, websites, and documentation, we can see that plain language is a dominant discursive goal today.[2]

Although our focus is on the American context, it is worth noting that plain English is pursued globally. Another U.S. government plain language mandate, the Securities and Exchange Commission's plain writing initiatives of 1998 and 2008 not only extend plain language into the private sector but also include the requirement that foreign

[1] For an example of the last, see the Department of Health and Human Services' 2021 Plain Writing Act Compliance Report.

[2] See, for example, the Microsoft Writing Style Guide and the Google Developer Documentation Style Guide.

Proceedings of Recent Advances in Natural Language Processing, pages 1180–1187
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_133

firms listing shares on U.S. stock exchanges use plain English in their prospectuses (SEC, 2021). There is also the Canada-based organization, Plain Language Association International (International, 2021), as well as one based in the UK, Clarity, which focuses on making legalese plainer (Clarity, 2021). Plain writing is nowhere—it is supposed to be inconspicuous writing that functions like a transparent medium for information—and everywhere.

Quantitative measures of plainness are generally out of favor these days among plain English advocates. Early on, the American plain language movement was associated with readability formulas: most notably the Flesch-Kincaid formula (still available in Microsoft Word), the similar Dale-Chall formula, and the Gunning fog index (Klare, 1963). But since the late 1970s, plain language advocates have adopted usability testing and emphasized design considerations beyond words and sentences: that is, information, document, and visual design (Redish, 2000; Schriver, 2017). We enumerate the main limitations of existing readability formulas below.

At the same time, there are quantifiable features of plain writing. The American government-advocacy network, the Plain Language Action and Information Network ("PLAIN"), includes among its techniques for writers the use of "short sentences" and "common, everyday words" (PLAIN, 2021). Similarly, the SEC's "Plain English Rule" is defined by six principles, the first two of which are "short sentences" and "definite, concrete, everyday language" (SEC, 1998b). And the *Oxford Guide to Plain English*'s first two guidelines that pertain to style—the guidelines immediately after "Plan before you you write" and "Organize your material..."—concern sentences and words: "Over the whole document, make the average sentence length 15-20 words," and "Use words your readers are likely to understand" (Cutts, 2020).

In this paper, we use quantitative measures of sentence length and word difficulty to evaluate some of the documents identified on the PLAIN website as good models of plain writing. But rather than combining sentence measures and word measures into a single readability score—a single score that is of little use to individual writers trying to make their writing plainer—we take the simple step of keeping each measure separate; we entertain the possibility that disjoined measures might be more illuminating and helpful to writers. We find, for example, that a text belonging to a domain (academic philosophy) oftentimes charged with jargon, does not use extremely difficult words, although it does have long sentences. Using these two separate measures, we reveal among PLAIN's exemplars a degree of variation that complicates current understandings of plain writing. Furthermore, in order to understand better what plainness is in all of its variety, we position PLAIN's different exemplary documents in relation to other documents written in other kinds of relatively accessible English (*The New York Times*, Voice of America Special English, and *Wikipedia*) and one academic document belonging to a genre perceived to be difficult (mentioned above). We propose that quantitative measures complement current approaches to plain writing and that advancing our knowledge about plain writing will require a better sense of the different types of plainer English.

2 The Problem with Existing Readability Measures

Plain language advocates have described several limitations to classic readability measures from the mid-twentieth century, including the Flesch-Kincaid formula. These measures, originally designed to measure children's reading abilities, do not accurately measure the reading abilities of adult information consumers. They also generate one-size-fits-all scores regardless of audience; worse, these scores do not typically give writers helpful guidance toward improving a piece of writing through revision. Most of all, for modern proponents of plain writing, readability formulas fail to take into account all of the non-prose elements of websites and other documents. These include the organization of information, the use of headings, tables of contents, layout and formatting, visuals, and so on (Redish, 2000; Redish and Selzer, 1985). Therefore, although readability measures have in the past been used to evaluate the accessibility of government documents, such measures are not mentioned in either the Plain Writing Act of 2010 or the SEC's Plain English Rule of 1998.

Existing quantitative approaches have additional limitations. They prove brittle in practice: most formulas measure a word's accessibility using the word's length in syllables or characters. The problem with this method is that many common words are long and many rare words are short: "international," "communication," "relationship," and "en-

tertainment" are far more likely to be understood than "mien," "feign," "pang," "dote," and "cinch."

The methods are also vulnerable to gaming (Redish, 2000). For example, Flesch-Kincaid readability scores can be raised by replacing lengthy words with acronyms. For example, an Internal Revenue Service (IRS) form designer describes replacing "self-employment tax" with "S.E. tax" in order to "improve" the document's readability score (National Public Radio, 2016).

More recent measures such as Lexile (Stenner, 1996), which use word frequency in reference corpora to measure word difficulty, solve some of the problems with earlier methods. But corpus-based measures of word difficulty bring with them new problems. One is the problem of estimating the difficulty of words that do not occur in the reference corpus.

A second problem is that Lexile and other formulas provide users with a single statistic. Quantitative measures of readability use two lists of numbers: the lengths of the document's m sentences $(l_1, l_2, \ldots, l_m)$ and the familiarity (or difficulty) of the document's n words $(r_1, r_2, \ldots, r_n)$. Flesch-Kincaid and Lexile scores are linear combinations of two statistics, one involving sentence lengths and another involving word difficulties.[3] But writers would benefit from finer-grained information than a summary score: whether their sentences could be more concise, or their vocabulary could be more commonplace, or both. It is for this reason that we disjoin the two components of popular readability formulas.

Third, today's measures penalize the judicious, infrequent use of technical terms. Because extremely rare words naturally occur in many genres, including documents that are required to be written in plain language, penalizing a document for having a few isolated rare words—as Lexile's averaging does—is unhelpful. For example, extremely rare words naturally occur in documents devoted to defining unfamiliar terms; countless plain language documents are devoted to this kind of explanatory, definitional work. An article written in plain language describing a coronavirus naturally uses the word "coronavirus." Such an article should not be penalized in proportion to the negative (log) frequency of the word. Indeed, such an article's use of the word should not be penalized *at all*.

Even when a document is not defining an unfamiliar term, penalizing a plain language document for an isolated rare word can be counterproductive. Technical terminology or the linguistic norm of "technicity" is not only an inevitable part of informational discourse, but oftentimes necessary for communicating ideas and communicating them comprehensibly (Guillory, 2004). Failing to mention that a technical term is often used to describe an item would be irresponsible since the reader may, in practice, only encounter the technical term. For example, a plain language description of how to ship goods overseas should be encouraged to mention that a list of goods for transport is called a "bill of lading," even though the word "lading" is spectacularly rare. To the extent that penalizing documents for exhibiting technicity encourages writers to avoid technical terms, received measures of plain language can inadvertently promote less comprehensible prose. In general, there is a strong case that none of the existing quantitative measures really encourages writing that is "plain" or easier to read.

3 Methods

We gather machine-readable versions of plain language exemplars featured on the US government's `plainlanguage.gov` website (maintained by PLAIN) as well as reference documents whose language is generally known (e.g., *New York Times* articles). For each document, we describe two empirical distributions: the distribution of sentence lengths and the distribution of word difficulties.

Note that we work with distributions and not summary statistics. Lexile, Flesch-Kincaid, and other familiar measures use averages of sentence- or word-level measurements of sentence complexity and vocabulary difficulty.

3.1 Features

Sentence lengths. We identify distinct sentences in machine-readable texts using a rule-based English language sentence tokenizer distributed with the NLTK software (Bird et al., 2009). We use the particular rule set which is distributed with version 3.5 of the software. These rules, derived from training on the WSJ portion of the Penn Treebank, have not changed since August 2013.

In order to arrive at a word count for each sen-

[3] A document's Lexile score, a measure of prose difficulty, is a rescaled version of $0.82247x - 2.14634y +$ constant, where x is the document's log mean sentence length and y is the mean log word accessibility, where accessibility is the frequency in a proprietary reference corpus (Stenner, 1996).

tence, we first tokenize the sentence using the Moses tokenizer (Koehn et al., 2007). We then remove all tokens that are not words. We define a word as a token which has characters in the following set: Unicode letters and the hyphen, with optional initial apostrophe (regular expression "`'?[\p{Letter}-]+`"). This definition is aligned with the Moses tokenizer, which preserves hyphenation and splits contractions. The number of tokens that remain after removing non-words is the sentence's length.[4]

There are, of course, other tokenizers and other methods for identifying distinct sentences. We use established methods to facilitate others reproducing our results.

Word difficulties. We follow the existing practice of measuring the accessibility of a word by how frequently it appears in a large reference corpus. To facilitate comparison we report all frequencies as frequencies per 1 billion tokens. To transform our measure to a measure of inaccessibility we multiply by -1.

For our reference corpus, we use the English language portion of the News Crawl corpus, published in association with the ACL's Third Conference on Machine Translation (WMT18). This corpus covers 11 years (2007-2017) and is inspired by and is a larger version of the "LM1B" language evaluation corpus (Chelba et al., 2014). After discarding duplicate sentences, we tokenize the corpus using the Moses tokenizer. This yields a corpus of 3.2 billion tokens (6.4 million types).

If a token is among the most common 100,000 types, we report its frequency per billion tokens as the measure of its accessibility. Otherwise, we estimate its frequency using regularized linear regression. Using such a model allows us to make serviceable estimates of the frequency of arbitrary tokens, including tokens which do not appear in the News Crawl corpus. Despite the corpus's size, countless technical terms are absent, as are neologisms introduced after 2017. This model takes as input the token's length in Unicode code points, its byte unigrams, byte bigrams, and byte trigrams. In calculating byte n-grams, we use UTF-8 encoding. The model outputs the token's estimated log frequency per 1 billion tokens. Additional details appear in Appendix A.

For the reasons described above, in this paper we avoid using summary statistics and report empirical distributions of these two features for each analyzed document.

3.2 Documents

Plain language exemplars The US government's website dedicated to the Plain Writing Act, `www.plainlanguage.gov`, offers the following documents as models of plain language documents. Given the context—a website designed to educate government officials on how to produce writing that conforms to the Plain Writing Act—we think it is appropriate to treat these documents as exemplars and not, say, marginal instances of documents conforming to the principles of the Act.

Several of the documents we use are available in the form of page images (PDFs). To reduce the labor required to transcribe text from page images, we randomly sample parts of documents. The specific sampling strategy is mentioned alongside the description of each document.

1. *The 9/11 Commission Report* by the National Commission on Terrorist Attacks (1,911 words). Published in 2004, the report describes events leading up to the September 11, 2001 attacks in the United States. We sample sections uniformly at random and collect paragraphs within each section.

2. *Draft Grazing Manual* by the Bureau of Land Management (915 words). Published in 1997, the section, "Range Improvements," is featured on the PLAIN website. It describes regulations concerning physical improvements to lands grazed by domestic livestock or wild animals. We use the entire section.

3. *National Park Service Museum Handbook, Part II* by the National Park Service (1,654 words). Published in 2000, the Handbook describes how to manage National Park Service museum collections. We randomly sample sections. The handbook features technical language specific to museum operations (e.g., "archival," "deaccessioning").

4. *Oak Ridge Reservation Annual Site Environmental Report* by the Department of Energy (1,654 words). Published in 2016, the 506-page report describes the results of environmental monitoring at the Oak Ridge Reservation (ORR) in Tennessee. The ORR hosts

[4] For an implementation of the Moses rule-based tokenizer, we use the `sacremoses` Python package.

facilities associated with the maintenance of US nuclear weapons. We sample sections at random.

5. *A Plain English Handbook* by the Securities and Exchange Commission (1,969 words). Published in 1998, the 83-page Handbook describes "well-established techniques for writing in plain English." The manual itself obviously uses the style and techniques it recommends, which is why we have included this document. We sample sections at random.

Reference documents

1. *Voice of America Special English* (2,243 words). Five articles randomly sampled from published articles on the Voice of America News in Special English website, https://learningenglish.voanews.com/. Texts written using VOA Special English, the most widely used successor of Basic English, use a vocabulary of about 1,500 words.

2. *New York Times* (1,783 words). A random sample of four Arts and Music section articles from *The New York Times*. Articles were truncated to 500 words. This sample is included as an example of writing addressed to general audience with considerable formal education.

3. *Wikipedia* (2,160 words). We gather paragraphs from five randomly sampled articles in the WikiText-2 corpus of "Good" and "Featured" articles (Merity et al., 2016). The articles selected are Xenon, USS Illinois, Mount Jackson, The Moth (TV episode), and Krak des Chevaliers.

4. *Academic philosophy* (2,025 words). We sample sections at random from *Bodies that Matter* (1993) by Judith Butler. We include this document as an example of non-plain writing. We considered several academic philosophy texts. Butler's text featured distinctly longer sentences.

Although some of the reference texts are aggregations of several documents, we refer to these aggregations as "documents."

4 Results

Figure 1 shows the distribution of sentence lengths and word difficulties for each analyzed document. All distributions exhibit positive skew. Sentence length distributions in the reference texts align well with prior expectations about document plainness. Word difficulty distributions in the reference texts are less distinctive but also roughly align with prior expectations. Plain language documents feature sentences which are shorter than those found in academic philosophy.

Sentence length and word difficulty distributions for the plain language exemplars vary with no consistent pattern. For example, the *National Park Service Museum Handbook* tends to use much shorter sentences than *Wikipedia* and the *New York Times*. At the same time, the *Handbook*'s words are not distinctly more accessible.

Two plain language exemplars, the *SEC Plain English Handbook* and the *Oak Ridge Environmental Report*, clearly differ in their use of short sentences and everyday vocabulary. 75% of sentences in the *SEC Handbook* use 20 words or fewer. In the *Oak Ridge Report*, only 52% of sentences have 20 words or fewer. The *SEC Handbook* uses much more accessible language. Ignoring instances of most common 500 words, 75% of words in the *Handbook* appear at a rate higher than 48,600 per billion tokens ($\log(48600) \approx 10.79$). (Familiar words occurring at this rate are "easily" and "require.") In the *Oak Ridge Report*, only 58% of words occur at similar rates.

5 Discussion

Our analysis indicates that writers needing to comply with the Plain Language Act can benefit from focusing on their sentences. With the exception of the *National Park Service Museum Handbook*, the documents that model plain writing according to PLAIN are less plain in terms of sentence length than our Wikipedia samples. Now it is possible that vocabulary simplicity causes longer sentences; this is typically true of writing in some controlled vocabulary languages, like Basic English, where sentences can run abnormally long (Igarashi, 2015). But we have found among our documents that the Voice of America Special English sample has the most commonplace words and shorter sentences, and the document with the longest sentences (the *Oak Ridge Reservation Annual Site Environmental Report* also uses the rarest words. A preliminary recommendation, then, is that government agencies aiming to write plainly use shorter sentences—an achievable goal. A future area of research would be

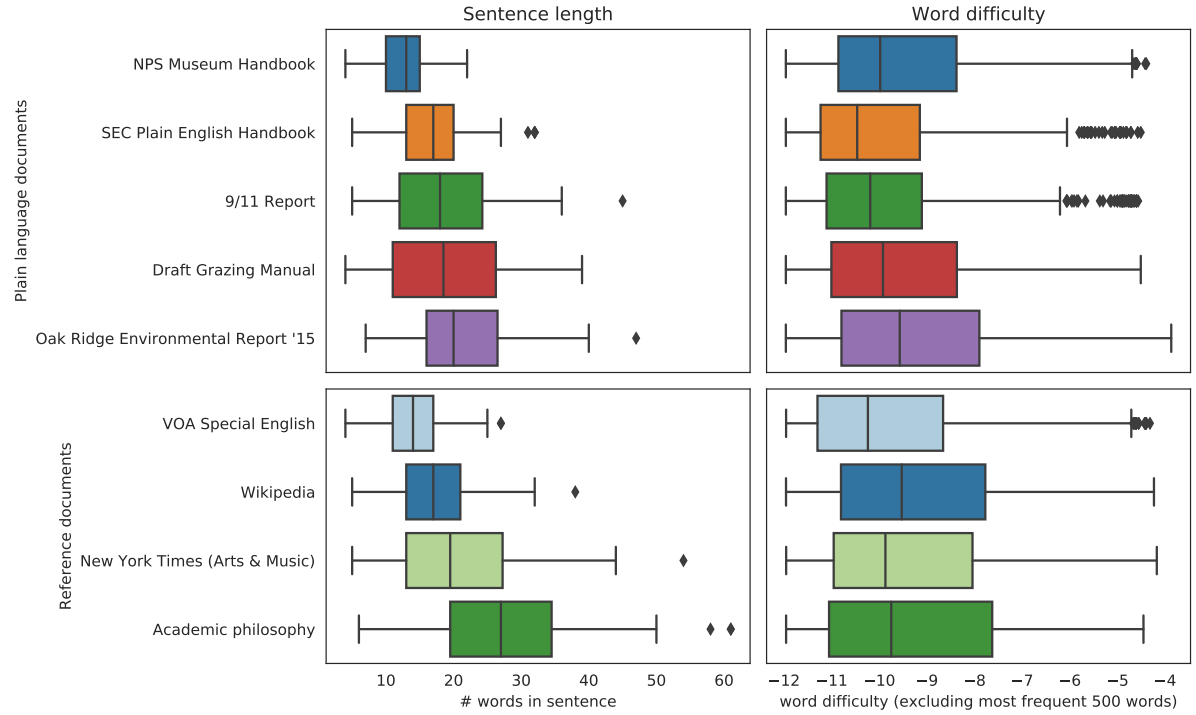

Figure 1: Sentence lengths and word difficulties. Word difficulty is defined as the negative log frequency in News Crawl corpus. Words among the most frequent 500 words are excluded in order to make the tails of the distributions visible.

to examine further the relationship between word rarity and sentence length with a larger sample of exemplary and reference documents.

In terms of vocabulary accessibility, the SEC Plain English Handbook is indeed exemplary, a valuable benchmark for other federal agencies striving to write in plain language. Its stylistic recommendations for translating abstract and obscure financial terminology form a helpful model for agencies writing about other subjects and domains (SEC, 1998a). Controlled vocabularies can also prove to be useful guides toward plainer writing: in particular, Voice of America's Special English strikes a good balance between vocabulary familiarity and sentence brevity. The style of the *Oak Ridge Reservation Annual Site Environmental Report* warrants reconsideration as an illustration of plain writing. We also hope to refine further our measure of word difficulty so that it is most useful for government employees.

A future line of inquiry would also consider how plainness manifests differently in different genres of informational writing. For example, do handbooks and manuals (e.g., the *National Park Service Museum Handbook* and the *SEC Plain English Handbook*) tend to exhibit briefer sentences than reports (e.g., the *Oak Ridge Reservation Annual Site Environmental Report*)? Our current findings are suggestive but not conclusive on this matter. But one hypothesis is that manuals and handbooks for practical purposes achieve sentence brevity more easily, whereas reports and other retrospective accounts have longer sentences due to these genres' goal of a comprehensive account.

Theoretical humanistic writing, although much maligned for the use of jargon and other difficult words (Culler and Lamb, 2003), also merits further investigation. Our sample of philosophical academese (Butler's *Bodies that Matter*) features rare terms less frequently than our *Wikipedia* sample and, surprisingly, at a similar rate as three plain writing exemplars (the *National Park Service Museum Handbook*, the *Draft Grazing Manual*, and the *Oak Ridge Reservation Annual Site Environmental Report*). According to our findings, Butler's writing is marked not by the use of jargon but rather by long sentences.

Finally, perhaps what we are dealing with is *plainer* writing rather than plain writing. Plainness is not a single, fixed quality possessed by any document but rather an ideal that different documents approach in various ways and with different

resulting textual features. Also, what seems unquestionably plain for one audience may not be plain for another. And, as we have seen, several documents deemed to represent plain writing are in fact quite variable in two of the enduring stylistic indicators of plainness, sentence length and word difficulty. Writing oriented to the plainness ideal and therefore made plainer than it would have otherwise been (hence all the before and after examples found in discussions of plain language) generates varieties of plainer writing.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. 1814425. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *arXiv:1312.3005 [cs]*.

Clarity. 2021. About Clarity.

Jonathan Culler and Kevin Lamb. 2003. Introduction: Dressing Up, Dressing Down. In Culler and Lamb, editors, *Just Being Difficult? Academic Writing in the Public Arena*, pages 1–12. Stanford Univ. Press.

Martin Cutts. 2020. *Oxford Guide to Plain English, Fifth Edition*. Oxford University Press.

John Guillory. 2004. The Memo and Modernity. *Critical Inquiry*, 31(1):108–132.

John Guillory. 2017. Mercury's Words: The End of Rhetoric and the Beginning of Prose. *Representations*, 138(1):59–86.

Yohei Igarashi. 2015. Statistical Analysis at the Birth of Close Reading. *New Literary History*, 46(3):485–504.

Plain Language Association International. 2021. Who We Are.

Julie Jennings and Jared C Nagel. 2020. Federal Workforce Statistics Sources: OPM and OMB.

George R. Klare. 1963. *The Measurement of Readability*. Iowa State University Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. *arXiv:1609.07843 [cs]*.

National Public Radio. 2016. How To Make Tax Forms Easier? Break The Math Up, One Step Per Line. https://www.npr.org/2016/03/31/472500987/how-to-make-tax-forms-easier-break-the-math-up-one-step-per-line. Online; accessed 3-July-2020.

PLAIN. 2021. What is plain language?

Janice C. Redish. 2000. Readability formulas have even more limitations than Klare discusses. *ACM Journal of Computer Documentation*, 24(3):132–137.

Janice C. Redish and Jack Selzer. 1985. The Place of Readability Formulas in Technical Communication. *Technical Communication*, 32(4):46–52.

Karen A. Schriver. 2017. Plain Language in the US Gains Momentum: 1940–2015. *IEEE Transactions on Professional Communication*, 60(4):343–383.

SEC. 1998a. A Plain English Handbook: How to Create Clear SEC Disclosure Documents.

SEC. 1998b. U.S. Securities and Exchange Commission plain writing rule - 421(d).

SEC. 2021. U.S. Securities and Exchange Commission plain writing initiative.

A. Jackson Stenner. 1996. Measuring Reading Comprehension with the Lexile Framework. In *North American Conference on Adolescent/Adult Literacy*, Washington, DC.

United States Congress. 2010. Plain Writing Act of 2010. Public Law 111-274.

A Estimating Word Frequencies

We follow the existing practice of measuring the accessibility of a word by how frequently it appears in published writing. Although this practice sounds easy to implement, doing so is complicated by the need to make estimates of the frequency of arbitrary words, including rare words, proper nouns, and neologisms. For high frequency words, estimates derived from frequencies in large corpora of

everyday texts (e.g., newspapers, magazines, general interest books) are serviceable. For many other words, this approach is not viable. Uncommon technical terms and proper nouns which appear in dictionaries are frequently absent from even the largest corpora. Neologisms and rare plural forms (e.g., *crowdfundings*, *virtuosas*) may not appear in dictionaries or large corpora but surely merit being assigned estimated frequencies higher than random character strings.

We solve this problem by using a simple model to estimate the frequency of uncommon words. We use regularized linear regression, also known as ridge regression, to predict a word's frequency per billion tokens. We extract the following features from the token: length in Unicode code points, byte unigrams, byte bigrams, and byte trigrams. The model is trained to predict the token's log frequency.

We train the model using token frequencies for all types which occur at least 50 times per billion tokens, reasoning that the News Crawl corpus contains a variety of incidental corruptions which we do not wish to model. We also exclude from the training data the most common 50,000 types, reasoning that the characters of extremely common words are not useful in predicting the frequency of rare words. We verify the model produces reasonable estimates by holding out 10% of the training data and asking the model to predict the log frequency of the held-out types.

The chief flaw with this particular approach is that it relies on a relatively homogeneous corpus of news articles. Words which tend to appear in news articles have inflated frequencies (e.g., *said*). Regularized linear regression also inflates the frequency of extraordinarily rare tokens (e.g., rare technical terms). Neither of these flaws is consequential in the present context. To study plain language we only need a general sense of how frequently a given word appears in everyday use.

Although the model is of token frequency, we only ever use frequencies of words. As described earlier in this paper, we define a word as a token which consists primarily of Unicode letters.